Expert Systems with Applications 42 (2015) 5848–5859

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Cluster analysis using optimization algorithms with newly designed objective functions

D. Binu

Aloy Labs, Bengaluru, India

ARTICLE INFO

Article history:
Available online 8 April 2015

Keywords:
Clustering
Optimization
Genetic algorithm (GA)
Cuckoo search (CS)
Particle swarm optimization (PSO)
Kernel space

ABSTRACT

Clustering finds various applications in the field of medical and telecommunication for unsupervised learning which is much required in expert system and its application. Various algorithms have been developed to clustering for the past fifty years after the introduction of k-means clustering. Recently, optimization algorithms are applied for clustering to find optimal clusters with the help of different objective functions. Accordingly, in this research, clustering is performed using three newly designed objective functions along with four existing objective functions with the help of optimization algorithms like, genetic algorithm, cuckoo search and particle swarm optimization algorithm. Here, three different objective functions are designed including the cumulative summation of fuzzy membership and distance value with normal data space, kernel space as well as multiple kernel space. In addition to the existing seven objective functions, totally, 21 different clustering algorithms are discussed and the performance is validated with 16 different datasets which are synthetic, small and large scale real data. The comparison is made with five different evaluation metrics to validate the effectiveness and efficiency. From the research outcome, the suggestion is presented to select a suitable algorithm among 21 algorithms for a particular data and results proved that the effectiveness of cluster analysis is mainly dependent on objective function and the efficiency of cluster analysis is based on search algorithm.

1. Introduction

Expert systems are intelligent software programs designed for taking useful and intelligent managerial decisions for various domains like, agriculture, finance, education, medicine to military science, process control, space technology and engineering. The expert systems require different data mining methods to support decision making process. Among different data mining methods, classification and clustering are two important methods applied widely for expert system. Clustering which is unsupervised learning has received significant attention among the researchers due to its wide applicability for the past fifty years after the introduction of k-means clustering algorithm (McQueen, 1967), which is well-known algorithm for clustering due to its simplicity. Due to the reception of k-mean clustering, variants of k-means clustering algorithms are introduced by different researchers by pointing out various problems like, initialization (Khan & Ahmad, 2004), k-value (Pham, Dimov, & Nguyen, 2004), and distance computation. One of the most accepted methods of clustering after the introduction of k-means clustering is fuzzy c-means clustering (FCM) (Bezdek, 1981), which is a popular algorithm after the k-means clustering, including fuzzy concept in computing the cluster centroids. FCM- algorithm is also found at various variants among the researchers (Ji, Pang, Zhou, Han, & Wang, 2012; Ji et al., 2012; Kannana, Ramathilagam, & Chung, 2012; Linda & Manic, 2012; Maji, 2011). The important variants of FCM algorithm is kernel fuzzy clustering (Zhang & Chen, 2004) and multiple kernel-based clustering algorithm (Chen, Chen, & Lu, 2011) which are based on the concept of FCM algorithm with the inclusion of kernels, accepted widely for its capability of doing the task for non-linear data. More interestingly, these entire algorithms have found importance in image segmentation (Chen et al., 2011; Zhang & Chen, 2004) and the relevant applications related to image segmentation (Ji, Pang, et al., 2012; Ji et al., 2012; Li & Qi, 2007; Sulaiman & Isa, 2010; Szilágyi, Szilágyi, Benyób, & Benyó, 2011; Zhao, Jiao, & Liu, 2013).

After the introduction of soft computing techniques, the clustering problem is transformed to optimization problem, finding the optimal clusters in the defined search space. Accordingly most of the optimization algorithms are applied to clustering problems. For example, the first and pioneer optimization algorithm called, GA (Mualik & Bandyopadhyay, 2002) is applied for clustering initially and then, PSO algorithm (Premalatha & Natarajan, 2008), Artificial Bee Colony (Zhang, Ouyang, & Ning, 2010), Bacterial Foraging Optimization (Wan, Li, Xiao, Wang, & Yang, 2012),

E-mail address: altimatebinu@gmail.com

http://dx.doi.org/10.1016/j.eswa.2015.03.031
0957-4174/

Simulated Annealing (Selim & Alsultan, 1991), Differential Evolution Algorithm (Das, Abraham, & Konar, 2008), and Evolutionary algorithm (Castellanos-Garzón & Diaz, 2013) and Firefly (Senthilnath, Omkar, & Mani, 2011) are subsequently applied for clustering. Recently, İnkaya, Kayalıgil, and Özdemirel, 2015 utilized Ant Colony Optimization for clustering methodology using two objective functions, namely adjusted compactness and relative separation. Liyong, Witold, Wei, Xiaodong, and Li (2014) utilized genetically guided alternating optimization for fuzzy c-means clustering. Here, interval number was introduced for attribute weighting in the weighted fuzzy c-means (WFCM) clustering to obtain appropriate weights more easily from the viewpoint of geometric probability. Hoang, Yadav, Kumar, and Panda (2014) have utilized the recent optimization algorithm called, Harmony Search Algorithm for clustering. Yuwono, Su, Moulton, and Nguyen (2014) have developed Rapid Centroid Estimation utilizing the rules of PSO algorithm to reduce the computational complexity and produced the clusters with higher purity. These recent algorithms utilized the traditional objective function for evaluating the clustering solution.

After that, hybrid algorithms are in the field of doing clustering process over the datasets to utilize the advantages of both the algorithms taken for hybridization. Here, two optimization algorithms are combined to do the clustering task as like, GA with PSO (Kuo, Syu, Chen, & Tien, 2012). From this, we can say that if any new optimization algorithms are being done, researchers are waiting to utilize the updating algorithm for clustering process. Due to the successful application of hybrid algorithms in clustering process, researchers are then hybridized the traditional clustering algorithms with the optimization algorithm. For example, GA is combined with k-means clustering, called genetic-k-means (Krishna & Murty, 1999) and the similar type of work is given in Niknam and Amiri (2010). Recently, Krishnasamy, Kulkarni, and Paramesran (2014) have proposed hybrid evolutionary data clustering algorithm referred to as K-MCI, whereby, K-means with modified cohort intelligence are combined for data clustering. Wei, Yingying, Soon Cheol, and Xuezhong (2015) have developed hybrid evolutionary computation approach utilizing Quantum-behaved particle swarm optimization for data clustering. Garcia-Piquer, Fornells, Bacardit, Orriols-Puig, and Golobardes (2014) have developed Multiobjective Clustering to guide the search following a cycle based on evolutionary algorithms. Tengke, Shengrui, Qingshan, and Huang (2014) have proposed a cascade optimization framework that combines the weighted conditional probability distribution (WCPD) and WFI models for data clustering.

In optimization-based clustering applications, clustering practices are operated, based on the fitness function, which validates the optimal cluster achieved. Here, the constraint is that fitness function developed should be capable of providing the good clusters' quality. The objective function is also responsible for the validation of the clustering output and directing it through the optimal cluster centroids. However, when looking into clustering' fitness functions, most of the optimization-based algorithm utilized the k-mean objective (minimum mean square distance) as fitness function for optimal searching of cluster task (Wan et al., 2012) because of its simplistic computation. Similarly, FCM objective is also applied as fitness function for finding the optimal cluster centroids (Ouadfel & Meshoul, 2012) due to its flexibility and its effectiveness. Also, authors utilize some cluster validity indices to apply on swarm intelligence-based optimization algorithm (Xu, Xu, & Wunsch, 2012) with the different perspective of cluster quality. In addition to, fuzzy cluster validity indices are developed with the inclusion of fuzzy theory and then, it is applied on optimization algorithm, GA (Pakhira, Bandyopadhyay, & Maulik, 2005). In a further way, the multiple objectives are combined to do the clustering optimization as like (Bandyopadhyay, 2011). Here, cluster stability and validity are combined as fitness and then it is solved using optimization algorithm, simulated annealing (Saha & Bandyopadhyay, 2009).

By means of the overall analysis, our finding is that most of the optimization algorithm utilizes k-means (KM) and FCM objective for their clustering optimization. Moreover with the best of our knowledge, MKFCM (Multiple Kernel FCM) objective is not solved previously through optimization clustering. So, with the intention of doing clustering task with optimization totally, two well known objectives (k-means and FCM), two recent objectives (KFCM and MKFCM) and three newly designed objective functions are utilized here. The reason of selecting these objectives is its (i) applicability and popularity (k-means and FCM objectives are chosen), (ii) regency and standard (KFCM and MKFCM are chosen), (iii) effectiveness and importance (three newly designed objective function). Then, we are in need of optimization algorithm for solving these objectives. Even though various optimization algorithms are presented in the literature, three optimization algorithms are chosen for our task of applying clustering process. Here, GA, PSO algorithm and CS algorithm are chosen because GA is traditional and popular one (Goldberg & David, 1989), PSO is an intelligent algorithm accepted by various researchers to its capability of changing the condition according to its most optimistic position (Kennedy & Eberhart, 1995), CS is a recent and effective algorithm proved better for various complex task of engineering problems (Yang & Deb, 2010).

The basic organization of the paper is given as follows: Section 2 provides contributions of the paper and Section 3 discusses objective measures taken from the literature. Section 4 presents new objective functions designed and Section 5 provides the solution encoding procedure. Section 6 discusses optimization algorithms taken for data clustering and Section 7 discusses the experimentation with detailed results. Finally, the conclusion is summed up in Section 8.

2. Contributions of the paper

The most important contributions of the paper are discussed as follows:

(i) *Clustering process with optimization:* We have developed clustering process through optimization technique in order to accomplish the optimal cluster quality. So, two traditional objective function (KM and FCM), two recent objective functions (KFCM and MKFCM) and three newly developed objective functions are operated to do the task. Moreover, the optimization algorithms such as, GA, PSO and CS algorithm are considered.

(ii) *Hybridization:* With the best of our knowledge, MKFCM objective is firstly solved with the optimization algorithms in this work. Hence, three optimization algorithms such as, GA, PSO and CS algorithm are combined with MKFCM objective functions to get three new hybridization algorithms, (GA-MKFCM, PSO-MKFCM, and CS-MKFCM) which are not presented in the literature previously.

(iii) *New objective functions:* We have designed three new objective functions (FCM + CF, KFCM + KCF, MKFCM + MKCF), including the cumulative summation of fuzzy membership and distance value. Here, the same cumulative summation is also performed with kernel space as well as multiple kernel space. Again, these three new objective functions are derived with good mathematical formulation and the corresponding theorem and the proof is moreover provided.